Lausunto kansallisesta big data -strategiasta

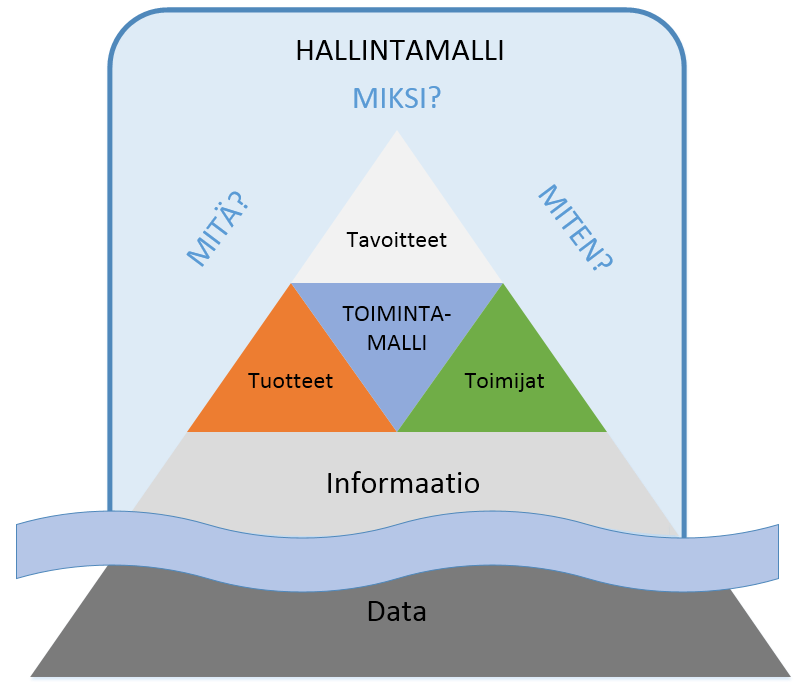
Lausunnon pyytäjä liikenne ja viestintäministeriö (LVM), joka toivoo erityisesti palautetta esitetyistä toimenpide-ehdotuksista. Lisäksi toivotaan tietoa siitä, minkälaisiin big dataan liittyviin toimiin Itä-Suomen yliopisto on ryhtynyt ja minkälainen rooli Itä-Suomen yliopistolla voisi kansallisessa big data -kehittämistyössä olla.

Itä-Suomen yliopistossa tehdään muun muassa seuraavia asioita big dataan liittyen:

* Ympäristöinformatiikka luo keinoja, joiden avulla ympäristöön liittyvistä suurista tietomassoista saadaan esiin olennainen ja tämä ydinasia voidaan esittää tiedon loppukäyttäjälle sopivimmassa muodossa[[1]](#footnote-1)
* Biolääketieteen[[2]](#footnote-2) opetuksessa ja tutkimuksessa hyödynnetään monipuolisesti big dataan liittyviä tekniikoita (ml. menetelmät ja työvälineet). Lisäksi Bioinformatiikkakeskus[[3]](#footnote-3) ylläpitää bioinformatiikan tutkimuksessa tarvittavaa infrastruktuuria.
* Fysiikan ja matematiikan[[4]](#footnote-4) sekä sovelletun fysiikan[[5]](#footnote-5) opetuksessa ja tutkimuksessa käytetään monipuolisesti big dataan liittyviä tekniikoita
* Tilastotieteessä annetaan koulutusta menetelmistä (kuten monimuuttujamenetelmät) ja työvälineistä (kuten R)[[6]](#footnote-6)
* Tietojenkäsittelytieteen tutkimuksessa, erityisesti älykäs medialaskenta[[7]](#footnote-7), hyödynnetään big data -tekniikoita ja tietojenkäsittelytieteessä on datatieteeseen[[8]](#footnote-8), väri- ja kuvateknologiaan[[9]](#footnote-9) sekä älykkääseen medialaskentaan[[10]](#footnote-10) painottuvat koulutuskokonaisuudet. Myös kansainvälisen IT-maisteriohjelman (IMPIT[[11]](#footnote-11)) yksi aihealue on big data.

Itä-Suomen yliopistolla voisi olla kansallisessa big data -kehittämistyössä seuraavia rooleja:

* osatoteuttaja kansallisissa ja kansainvälisissä hankkeissa, joissa tarvitaan big data -osaamista
* puolueeton tekniikka-arvioija, kun arvioidaan ohjaamattomien (unsupervised) ja ohjattujen (supervised) menetelmien ja niitä tukevien työvälineiden sopivuutta big datasta löytyvien muuttujien riippuvuuksien arvioinnissa ja edelleen tulevaisuuden ennakoinnissa (prediction) ja toiminnan ohjaamisessa (prescription)
* big data ekosysteemin tarjoaja opetukseen ja tutkimukseen. Big data -asioita tulee tarkastella organisaatiotasolla (julkisyhteisöt ja yritykset) osana kokonaisuutta eli mitä dataa hyödynnetään, miksi hyödynnetään ja miten hyödynnetään (Kuva 1).



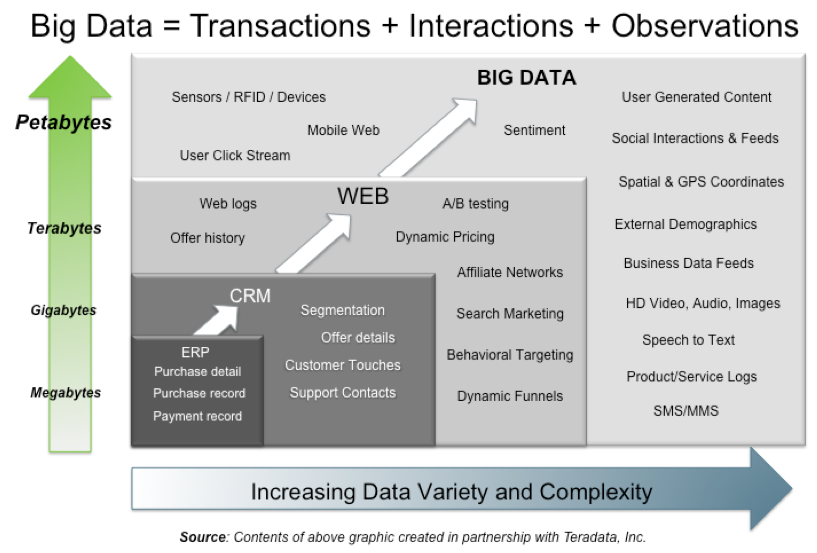
Kuva . Organisaatio omien ja ulkopuolisten datojen hyödyntäminen on osa organisaation toimintaa[[12]](#footnote-12)

Itä-Suomen yliopisto lausuu seuraavaa toimenpide-ehdotuksista (kaikista puuttuu vastuutaho, konkreettiset tavoitteet, aikataulut, mittarit):

1. ”Datalähettiläitä” hankitaan tuomaan big data- tietoutta ja osaamista yritysten ja hallinnon käyttöön toteutettuna mm. kansainvälisinä vierailijaluentoina ja osaajavaihtona
   * myös kansallisella tasolla on löydettävissä ”datalähettiläitä”, joita kannattaa hyödyntää
   * nimiehdotus PhD Vincent Granville
   * ehdotus vierailijaluennoitsijoiden esitykset videoidaan, jotta niitä voidaan katsoa, vaikka ei pääsisi osallistumaan tilaisuuksiin
   * osaajavaihtoon kannattaa koota kansallisella tasolla halukkaat ja järjestelmällisesti koordinoida osaajavaihtoa
2. Luodaan alan yrityksiä, tutkijoita ja asiantuntijoita kokoava avoin Big data-klusteri sekä osaamisen kehittämistä tukevaa verkostotoimintaa tämän ympärille (Tekes, Teknologiateollisuus, Tekniikan Akateemiset, Ohjelmistoyrittäjät, TTL, yritykset jne).
   * loistava ehdotus
   * virtuaaliyhteisö kuten LinkedIn-tyyppinen kanava ajatusten vaihtoon sekä informaation jakamiseen
3. Julkisen hallinnon organisaatioihin nimetään datavastaavia, joiden tehtäviin kuuluu datan keruun ja analysoinnin järjestäminen. Yksityisiä yrityksiä ja muita yhteisöjä varten voidaan järjestää neuvontaa.
   * useimmilla julkisyhteisöillä on tietosuojavastaavat ja osassa julkisyhteisöistä he voinevat toimia myös datavastaavina
   * datan keruun ja analysoinnin järjestäminen edellyttää tietoa datatarpeista, jotka tulee tunnistaa esimerkiksi kokonaisarkkitehtuurityön yhteydessä
4. Vaikutetaan siihen, että peruskoulun opetussuunnitelmiin sisältyy dataan käyttöön liittyvää opetusta vuonna 2016 ja tuetaan open source teknologioiden leviämistä.
   * tulee määritellä tarkemmin, mitä datan käytön opetus tarkoittaa
   * lisäksi olisi hyvä määritellä, mitä OS-teknologioita käytetään
5. Kartoitetaan opetushallinnon johdolla tarvittavia osaamisprofiileja ja olemassa olevaa koulutustarjontaa.
   * ensin olisi hyvä tunnistaa itse osaamiset ja sen jälkeen niitä voidaan allokoida eri rooleille
6. Koordinoidaan kansallisesti käynnisteillä ja sunnitteilla olevia big data koulutusohjelmia.
   * jo käynnissä olevista tulee kerätä kokemukset ja varmistaa, että ne jatkuvat
7. Osallistutaan EU-laajuisen big dataan erikoistuvien yliopistojen verkoston luomiseen.
   * loistava ehdotus
   * kannattaa vahvistaa myös kansallista verkostoa
8. Ohjelmointi/tietojenkäsittely/bisnesprosessien opiskelijoita kannustetaan yrityksiin big data harjoitteluun, projekteihin ja opinnäytetöihin (hyvänä esimerkkinä Aallossa diplomityöparit)
   * mielellään ei vain yrityksiin, vaan yrityksiin ja julkisyhteisöihin
   * harjoittelutukirahoitus ja mahdollisesti kansallinen koordinointi edesauttaisi asiaa
   * Itä-Suomen yliopiston Tietävä-projektissa yritykset ja julkisyhteisöt ovat antaneet esimerkkiaineistoja analysoitavaksi ja Tietävä-tiimi on tehnyt aktiivista yhteistyötä organisaatioiden kanssa sekä auttanut organisaatioita löytämään heille sopivan harjoittelijan, projekityöntekijän tai tutkielman tekijän
9. Toteutetaan reaaliaikaista tarvekartoitusta siitä, kuinka paljon ja millaista osaamista yritykset tarvitsevat (big data yritysklusteri/foorumi)
   * mielellään ei vain yritykset, vaan yritykset ja julkisyhteisöt
10. Täydennys- ja muuntokoulutusta uudistetaan vastaamaan big data osaamistarpeita
    * ensin ne osaamistarpeet eksplisiittisesti esiin ja sitten kootaan koulutusta ensisijaisesti olemassa olevasta tarjoomasta – pyörää ei kannata keksiä kerta toisensa jälkeen uudelleen
    * kansallisella tasolla pitää saada ehdottomasti MOOC-kursseja, joissa on laadukkaat videot ja tehtävät – tarvittaessa kursseihin voidaan yhdistää lähipäivä, joissa ollaan sormet näppäimillä eli tehdään konkreettisesti data-analytiikkaa
11. Tuetaan yritysten big data osaamista kehittävien sisäisten hautomojen toimintaa, osaamisvaihtoa ja työnkiertoa
    * mielellään ei vain yritykset, vaan yritykset ja julkisyhteisöt
    * onnistuisikohan osaamisenvaihto opinahjojen kesken, puhumattakaan työnkierrosta?
12. Kehitetään kansallista big data infrastuktuuria verkostomaisella yhteistyöllä
13. Tunnistetaan keskeiset standardointiprosessit joissa tulee olla mukana sekä organisoidutaan toimimaan näissä
14. Standardien luominen erityisesti datan varastointiin (storage) ja siirtoon (exchange) kehittää datan käsittelyn tekniikoita, työskentelymetodeja ja tehokkaita algoritmeja.
    * miksi?
15. Tietosuoja ja yksityisyys on myös teknologinen kehityskysymys. Panostetaan tieto/yksityisyydensuojan ratkaisujen teknologiakehitykseen (esim. Tekes)
    * ensin pitänee saada vastaukset kysymyksiin mitä suojataan ja miksi suojataan sekä mitä esimerkiksi uusi EU:n tietosuoja-asetus vaatii
16. Datan hyödyntämiseen ratkaisuja etsiviä ja kehittäviä henkilöitä rekrytoidaan hallintoon, esimerkiksi kumppanikoodarimallin mukaisesti, College-to-govt harjoitteluohjelmat tms.
17. Yhteentoimivuuden kehittäminen datan jakamisen standardeilla ja yhteishankkeilla erityisesti kunta- ja kaupunkiympäristössä (mm Kuusaika- hankkeen puitteisa)
18. Määritellään kansallisen dataportaalin hyödyntäminen myös big data kanavana Avoimen tiedon ohjelman osana (VM)
    * miksi [www.hri.fi](http://www.hri.fi) rinnalle on pitänyt tehdä <http://beta.avoindata.fi/fi>?
19. Velvoitetaan virastoja tunnistamaan data-aineistoja, joista big data kehityksen kannalta saataisiin merkitäviä hyötyjä. Tuetaan tässä virastojen pilottiprojekteja (Avoimen tiedon ohjelma/kehykset).
20. Tehdään selvitys big datan käyttöön vaikuttavista laeista
    * mitä big datan käytöllä tarkoitetaan?
21. Varmistetaan, että EU:n tietosuoja-asetus ja sen soveltaminen Suomessa mahdollistaa big data kehityksen yksityisyydensuojaa vaarantamatta
    * muun muassa EU:n tietosuoja-asetus kieltänee profiloinnin, jota big dataa analysoimalla tehdään
    * big data -kontekstissa jouduttanee tekemisiin muun muassa suostumusten hallinnan osa-alueella
22. Kehitetään suomalainen Big data "etiketti" ohjesääntö
    * suosittelisin periaatteita, sillä niihin liittyen annetaan myös toimintaohjeita/-sääntöjä
23. Selvitetään tarvetta Big data-hubille, erityisesti isojen ja pk-yritysten yhteisille datahankkeille ja datan vaihdolle liiketoiminnassa. (DIGILE, Teknologiateollisuus, yritykset)
    * teknologia mahdollistaa
24. Edistetään erityisesti pk-yritysten ja start-uppien big data resursseja (välineitä ja pääsyä aineistoihin) ja osaamista. Aktivoidaan näitä yrityksiä myös mukaan kansallisiin hankkeisiin ja yritysten aloitteisiin (vrt. GE:n tarjoamat fasiliteetit)
    * ”osaamisen siirto” tapahtuu parhaiten yhdessä tekemällä eri työmuotoja käyttäen. Lisäksi pelkkä aineisto ei paljon auta, jos ei ole harmaata aavistustakaan, mitä sille tekisi ja miksi (puhumattakaan miten).
25. Luodaan kansallinen data-analytiikan toimijoiden verkosto (tutkimuslaitos/yliopistoyhteistyö), jossa edistetään myös teknologiasiirtoa yrityksiin (menetelmät, työkalut, osaaminen)
    * loistava ehdotus
26. Osallistutaan ja jaetaan kansallisesti näkemyksiä EU:n Big/NESSI data foorumin toimiin
    * <http://www.big-project.eu/> oli uusi tuttavuus, jota kannattanee seurata LinkedIn:ssä
    * <http://www.nessi-europe.com/default.aspx?Page=home> oli hieman outo suositus, miksi ko foorumin toimintaan tulisi osallistua?
    * on olemassa joukko muitakin foorumeita, joihin kannattaa osallistua tai joita kannattaa ainakin seurata kuten <http://www.datasciencecentral.com/> - ko foorumit kannattaa koota ja perustella, miksi ko foorumi on tärkeä
27. Kehitetään joukkoistamiseen perustuvia big data- malleja, joista parhaillaan syntyy uudentyyppisiä laajoja ekosysteemejä.
    * ”teknologisen kehityksen myötä ammattilaisten ja harrastelijoiden kuilu on kaventunut. Joukkoistaminen on liitetty hyvin usein toisiin samankaltaisiin ilmiöihin, kuten avoimeen innovaatioon ja avoimen lähdekoodin hankkeisiin”[[13]](#footnote-13)
    * mitä joukkoistamiseen perustuvat big data -mallit ovat?
    * mitkä valitaan ja millä perusteella kehitettäväksi?
    * kuka ja ketkä kehittää?
28. Kansallista tukea ja verkostoa EU-tukiohjelmien hyödyntämiseen vahvistetaan erityisesti big data painotuksella (Horizon verkosto, Tekes, kehittäjäorganisaatiot)
    * loistavaa – kansallinen vastuutaho ja henkilö(t)?
29. Tuetaan erilaisia big data kokeiluja ja jaetaan näistä saatuja kokemuksia. Erityisen tarpeelliseksi katsotaan datan saatavuus hallituissa kokeiluissa, joihin liittyy esimerkiksi mobiilipaikantaminen
    * kansallinen vastuutaho ja henkilö(t)?
30. Disruptio datan vaikutuksesta tulevat voimakkaasti myös liiketoimintamalleihin: kokeilevat kehityshankkeet hyödyn ja tulonjaonmallien kehittämiseksi (vrt. myös yritysten yhteistyön aktivointitoimet)
31. Pääomaehtoisen rahoituksen (yksityiset ja julkiset pääomasijoittajat) saatavuutta tulee edistää
    * kansallinen vastuutaho ja henkilö(t)?
32. MyData selvityksen malleja testataan valituilla alueilla kehityksen vauhdittamiseksi ja uusien tietomallien yleistymiseksi eri alojen yritysten avulla. Pitkällä aikavälillä luodaan alustaa MyDatalle, johon yksilö voi omaa tietoaan kerätä ja haluamallaan tavalla jakaa sekä hyödyntää.
33. Yhteistyö Iso-Britannian MiData-labin kanssa Midata-kehityksen vahvistamiseksi

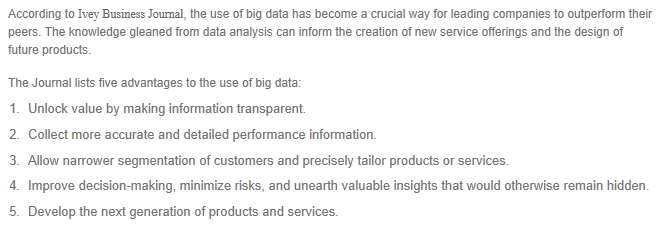
Muuta huomioitavaa:

* *1.1 Big datan aikakausi*. Big datan syntymiseen liittyen voisi ottaa people-to-people-, people-to-machine- ja machine-to-machine-näkökulmat[[14]](#footnote-14) ja/tai havainnollistaa big dataa usein käytetyllä kuvalla (Kuva 2), jossa Big data on tapahtumadata (transactions) + vuorovaikutusdata (interactions) ja havaintodata (observations).

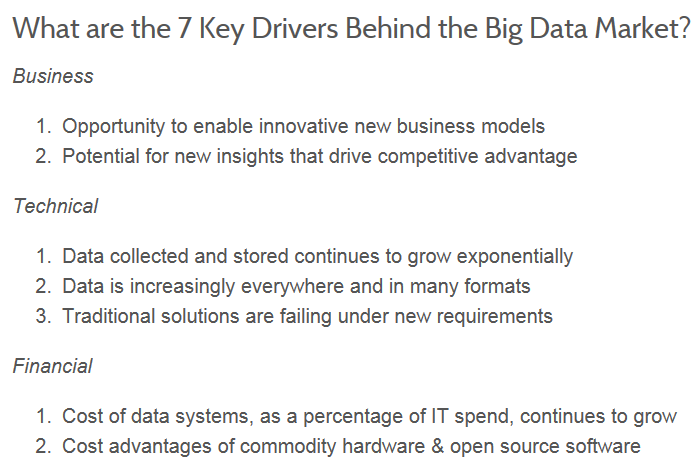


Kuva . Big datan muodostumisen havainnollistaminen[[15]](#footnote-15)

* Luvussa 1.1 voisi tuoda selkeästi esille sen, miksi big datasta ollaan kiinnostuneista (Kuva 3 ja Kuva 4).



Kuva . Big data -hyötyjä[[16]](#footnote-16)



Kuva . Big data -ajureista[[17]](#footnote-17)

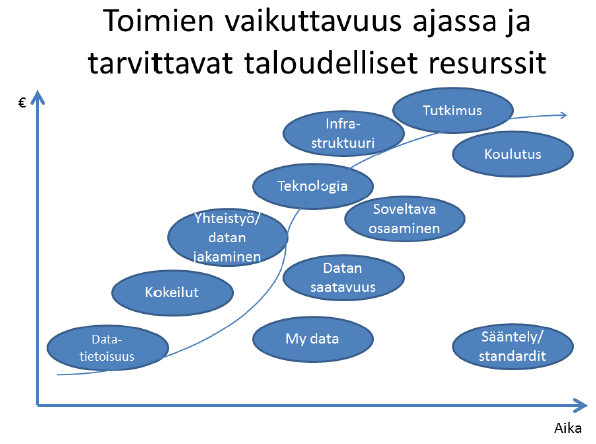
* Big data -strategian viimeistelyssä voi käyttää mainostoimistoa tai etsiä työryhmästä henkilö(t), jotka osaavat sekä visualisoinnin että tarinan kirjoittamisen – strategia voi olla narratiivi (tarina), joka rakentuu jonkin peruskuvan (kuten Kuva 5) ympärille, jota täydennetään ja josta korostetaan osia narratiivin eri vaiheissa. Lisäksi kannattanee havainnollistaa eri datalähteistä saatavan datan käyttöä (Kuva 6) eli vastata kysymykseen miksi. 

Kuva . Big datan syntylähteet ja hieman käsittelyprosessia[[18]](#footnote-18)

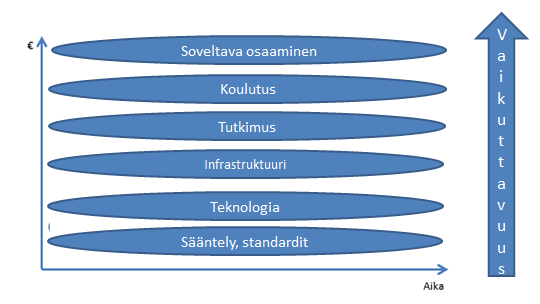


Kuva . Big datan käyttötarkoituksien havainnollistaminen[[19]](#footnote-19)

* *1.1 Big datan aikakausi* - ”Syntyvä data on muodoltaan hyvinkin erilaista”. Ilmeisesti tarkoitetaan erilaisia dataformaatteja, joista käytetään kansallisella tasolla myös tiedostomuoto-termiä[[20]](#footnote-20).
* *2.4 Tutkimus* - ”Big data-menetelmien käyttö on suomalaisessa tutkimuksessa vielä jokseenkin vähäistä”. Strategiassa tulee kirjoittaa auki, mitä big data –menetelmillä tarkoitetaan. Jos menetelmiin kuuluu esimerkiksi ohjaamattomat ja ohjatut koneoppimisen menetelmät, niin niitä on suomalaisessa tutkimuksessa käytetty – kansainvälisesti merkittävää tutkimusta tekee esimerkiksi professori Martti Juhola Tampereen yliopistolla. Myös big data -menetelmiin kuuluviin algoritmeihin ja operaatioanalyysimenetelmiin liittyvää tutkimusta tehdään kansallisella ja kansainvälisellä tasolla laajamittaisesti. Lisäksi eri sovellusalueilla, kuten bioinformatiikka ja ympäristöinformatiikka, tehdään ja on tehty merkittävää kansallista ja kansainvälistä tutkimusta.
* *2.5 Infrastruktuuri* – ”Oleellisessa osassa ovat myös erilaiset yhteentoimivuutta varmistavat toimet”. Mitä yhteentoimivuutta varmistavilla toimilla tarkoitetaan tässä kontekstissa?
* *2.6 Tekniset käytännöt ja standardit* – Big datan myötä data-analytiikka kehittyy huimaa vauhtia ja erityisesti ennakoivassa analytiikassa on havaittavissa erilaisia trendejä[[21]](#footnote-21) - yksi trendeistä on PMML:n[[22]](#footnote-22) käyttö. Predictive Model Markup Language (PMML) on de facto -standardi, joka mahdollistaa mallien kehittämisen työkaluriippumattomasti[[23]](#footnote-23).
* *7. Vaikuttavuus*. ”Pylpyrät” antavat sellaisen kuvan (Kuva 7), että big dataan liittyen ollaan ottamassa vasta ensiaskeleita – kuvasta saa sellaisen käsityksen, että big data asioissa ollaan vasta alkumetreillä ja se ei pidä paikkaansa. Soveltavan osaamisen -pylpyrään voisi yhdistää datatietoisuuden ja kokeilut (Kuva 8). Infrastruktuuri-pylpyrään voisi yhdistää yhteistyö/data jakaminen ja datan saatavuuden. Lisäksi sääntelyllä, kuten tulevalla EU:n tietosuoja-asetuksella, tulee olemaan merkittävä vaikutus big data ekosysteemeihin ja ne tulee ottaa huomioon heti alkuvaiheessa.

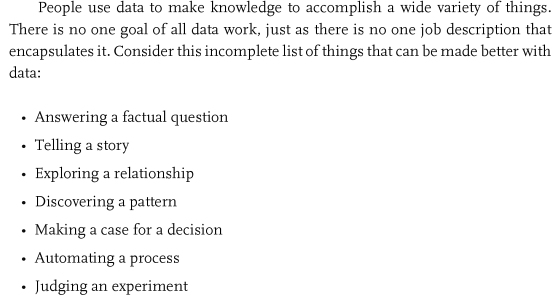


Kuva . sivun 45 kuva



Kuva . Ehdotus uudeksi kuvaksi

* *7. Vaikuttavuus* - ”Datan arvoketju ja arvonluonti A) Datan arvoketju: 1. data hankinta (acquisition) 2. analyysi 3. tuki ja joukkoistaminen (curation) 4. varastointi 5. käyttö B) Arvonluonti: miten datasta saadaan arvoa (data extraction)”. Kannattanee lähteä liikkeelle siitä, että arvioidaan, mitä dataa tarvitaan ja miksi (assessing the data). Jos tarvittava data on ns. raakadataa, niin se pitää joko itse ”murskata” (crunching the data) tai hankkia (acquisition) se ”kaiveluvalmiina” eli raakadatasta on valittu ominaisuudet, joita halutaan tarkastella (feature extraction). Datan ennakkoluulotonta tarkastelua voisi kutsua datan ”kaiveluksi” (digging into the data) ja loppupää voisi olla informaation kulutusta (information consumption) eli datapohjaisesti voimme tehdä (Kuva 9) esimerkiksi päätöksiä. Itse asiassa datan hyödyntäminen voidaan kuvata yksikertaisella prosessilla



Kuva 9. Shron Max (2014) Thinking with Data: How to Turn Information Into Insights[[24]](#footnote-24)

* *8. Strategian toimeenpano* - ”Luodaan yhteistyössä tiekartta toimien vastuista, etenemisjärjestyksestä, aikatauluista, vaikuttavuudesta ja mahdollisista riskeistä. Vahvistetaan strategian toimeenpano valtioneuvoston periaatepäätöksellä ja/tai hallitusohjelmassa. Toimet, jotka voidaan käynnistää heti resursoidaan ja organisoidaan välittömästi.”. Kansallisella tasolla tulee löytää toimija, joka on vastuussa ainakin julkisyhteisöjen big dataan liittyvistä -palveluista. Esimerkiksi tällainen toimija voisi olla Valtion tieto- ja viestintätekniikkakeskus Valtori[[25]](#footnote-25)

1. <http://www.uef.fi/en/envi/home> [↑](#footnote-ref-1)
2. <http://www.uef.fi/fi/biolaaketiede> [↑](#footnote-ref-2)
3. <http://www.uef.fi/fi/bioinformatics/bioinformatics-center> [↑](#footnote-ref-3)
4. <http://www.uef.fi/fi/fysmat> [↑](#footnote-ref-4)
5. <http://www.uef.fi/fi/sovfys/tutkimus> [↑](#footnote-ref-5)
6. <https://www.uef.fi/fi/cs/tilastotiede> [↑](#footnote-ref-6)
7. <http://www.uef.fi/fi/cs/tutkimus> [↑](#footnote-ref-7)
8. <http://www.uef.fi/fi/cs/datatiede> [↑](#footnote-ref-8)
9. <http://www.uef.fi/fi/cs/vari-ja-kuvateknologia> [↑](#footnote-ref-9)
10. <http://www.uef.fi/fi/cs/alykas-medialaskenta> [↑](#footnote-ref-10)
11. <http://cs.joensuu.fi/impit/> [↑](#footnote-ref-11)
12. kuva on [↑](#footnote-ref-12)
13. <http://fi.wikipedia.org/wiki/Joukkouttaminen> [↑](#footnote-ref-13)
14. <http://www.alphasixcorp.com/images/big-data-infograph.jpg> [↑](#footnote-ref-14)
15. <http://hortonworks.com/blog/7-key-drivers-for-the-big-data-market/> [↑](#footnote-ref-15)
16. <http://predictivescience.com/big-data-revolution-brings-disruption-innovation-and-opportunity/> [↑](#footnote-ref-16)
17. <http://hortonworks.com/blog/7-key-drivers-for-the-big-data-market/> [↑](#footnote-ref-17)
18. <http://twirl-project.eu/wp-content/uploads/2014/04/homePage.png> [↑](#footnote-ref-18)
19. <http://m.c.lnkd.licdn.com/mpr/mpr/p/2/005/05f/067/181883a.jpg> [↑](#footnote-ref-19)
20. [http://www.hri.fi/fi/data-haku/#](http://www.hri.fi/fi/data-haku/) [↑](#footnote-ref-20)
21. <http://www.predictive-analytics.info/> [↑](#footnote-ref-21)
22. <http://www.dmg.org/> [↑](#footnote-ref-22)
23. <http://www.dmg.org/products.html> [↑](#footnote-ref-23)
24. <http://www.google.fi/books?hl=fi&lr=&id=tqidAgAAQBAJ&oi=fnd&pg=PR4&dq=%22digging+into+the+data%22&ots=9CQEuzNgLv&sig=HLd7k0njvTcA7mQbHhd-oGoTYAU&redir_esc=y#v=onepage&q=%22digging%20into%20the%20data%22&f=false> [↑](#footnote-ref-24)
25. <http://www.valtori.fi/fi-FI> [↑](#footnote-ref-25)